

# Project Goals

## 2019 Chicago West Nile Virus Action Plan

**Team:**

Andrew Cooper  
Rachel Dudle  
Stephen Hage

Mike Kapelinski  
Ted Inciong  
Rahul Sangole

# Table of Contents

---

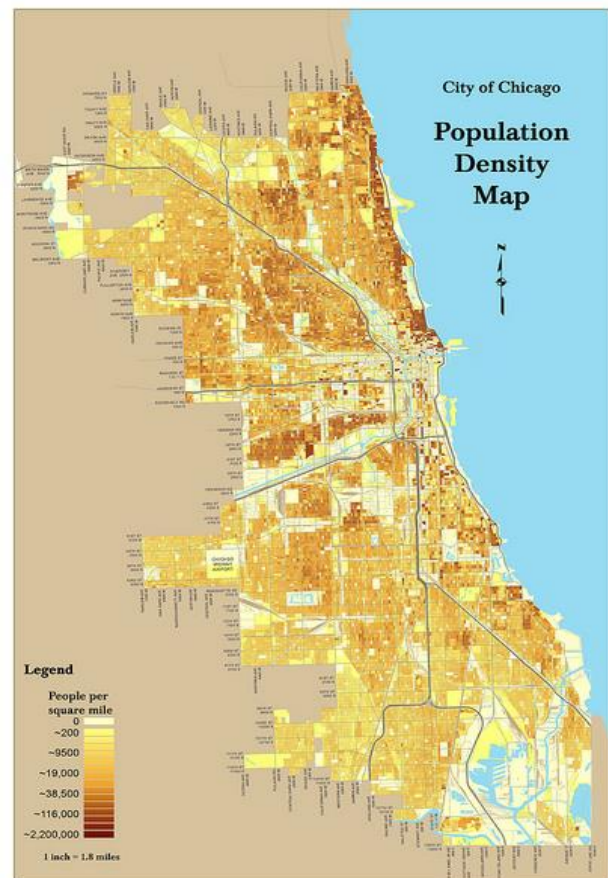
Overview.....	2
Business Case.....	3
Goals and Objectives .....	4
Data Source.....	5
Deliverables.....	7
Dashboard Visualization.....	7
App Development.....	7
Predictive Modeling.....	8
Risks.....	9
Software and Analytics Tools.....	9
Project Plan.....	10
Team.....	11

# Overview

In September 2001, West Nile virus was first identified in Illinois when laboratory tests confirmed its presence in two dead crows discovered in the Chicago area. This comes only two years after West Nile virus first emerged in the United States in New York in the fall of 1999. By the end of 2002, Illinois had counted more human cases (884) and deaths (64) than any other state in the United States<sup>1</sup>.

This is where SMARRT consulting group can help to prevent this costly pandemic from resurfacing and preserve public safety. SMARRT Analytics has focused exclusively on consultation in matters of public health to assess the risk of disease outbreaks in cities across the world to provide analytical expertise ultimately providing recommendations for intervention and prescriptive prevention.

The project outlined below proposes means to assess and provide analytical insights to prevent a West Nile outbreak in the Chicago area. The goal is to assess Chicago for West Nile prevalence. We will also establish the most influential factors which may contribute to an outbreak. We can then predict the potential of an outbreak and the best mitigation and prevention program for a city of 2.7 million residents.



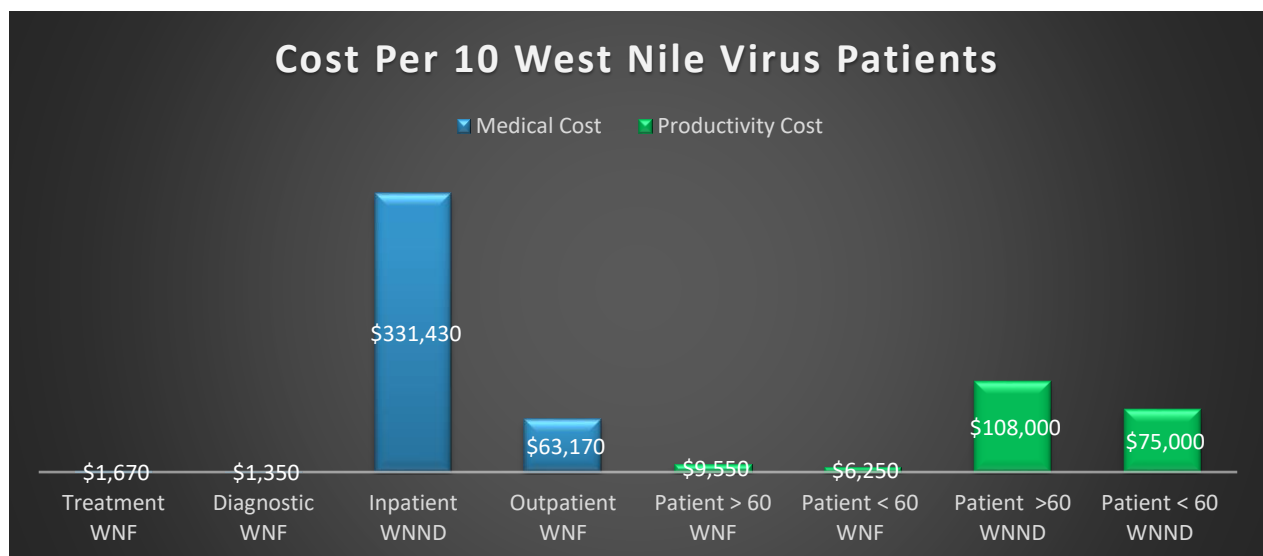
<sup>1</sup> <http://www.dph.illinois.gov/topics-services/diseases-and-conditions/west-nile-virus>

## Business Case

In 2002, West Nile virus was discovered in Chicago for the first time with over 225 cases reported. In response, the Chicago Department of Public Health (CDPH) has implemented a city-wide surveillance and mosquito control measure program. Based upon a 2010 study produced by the CDC, a 2005 outbreak of West Nile virus cost Sacramento County, California \$2.98 million. West Nile Virus (WNV) can have two different effects on a human host, West Nile Fever (WNF) and it's much more severe and costly West Nile neuroinvasive disease (WNND). A cost benefit analysis was also performed during this study, which indicated that preventative measures such as spraying would only need to prevent 15 cases of WNND to make the control measure cost effective.<sup>3</sup> Today the population of Chicago is 2.7 million which is 1.8 times the size of Sacramento County, California and the outbreak we experienced in 2002 having 225 infected compared to the 163 cases in California would result in an overall cost of over \$4 million in medical and productivity costs alone.

The CDPH recently commissioned a study which revealed the city of Chicago can significantly reduce costs associated with treating West Nile virus in hospitals and clinics through simple preventative measures. These include but are not limited to community level control programs such as targeted mosquito spraying of high-risk area and removal of debris associated with mosquito breeding (i.e. areas with stagnant water). In addition, personal protective measures such as use of mosquito repellent and wearing long sleeves have shown to decrease exposure to infected mosquitoes.

Therefore, the CDPH opened a requisition to modernize their mosquito controls system. The CDPH would like a system to identify top and emerging high-risk areas for the West Nile virus as well as a dashboard which allows users to monitor the West Nile virus in real-time. This modernized system will allow the CDPH to prevent the spread of West Nile virus.



<sup>2</sup>Healthy Chicago Data Brief - West Nile Virus. (n.d.). Retrieved from <https://www.chicago.gov/city/en/depts/cdph.html>

<sup>3</sup>Economic Cost Analysis of West Nile Virus Outbreak, Sacramento County, California, USA, 2005 - Volume 16, Number 3-March 2010 - Emerging Infectious Diseases journal - CDC. (2010, December 14). Retrieved from [https://wwwnc.cdc.gov/eid/article/16/3/09-0667\\_article](https://wwwnc.cdc.gov/eid/article/16/3/09-0667_article)

## Goals

SMARRT Consulting Group will help Chicago decrease the amount of West Nile virus infections in a cost-effective way. In order to accomplish this goal, guidance will be given to the city where preventative measures (like spraying) would be most effective, and where the most vulnerable populations live. We recognize that Chicago Department of Public Health already undertakes mosquito abatement via screening, targeting areas based on mosquito trap test results. These abatement efforts curtail mosquito population growth and reduce transmission of West Nile virus but cannot eliminate it. SMARRT Consulting Group's advanced models can improve upon existing methods for targeting mosquito spraying by identifying times and places where risk remains high and identification of risk areas early can prevent mosquito problems from increasing. In addition to using advanced predictive modeling techniques, SMARRT will use ancillary data sources that will better identify areas where neighborhood characteristics contribute to mosquito growth.

In addition to refining predictive models to identify areas to target with spraying, SMARRT will identify high risk regions where there is increased risk for human transmission and neuroinvasive disease. We will identify areas with a high concentration of vulnerable people using demographic data, school and senior center locations and other data sources.

This effort would also alert consumers as to the risk of West Nile virus in their location, so the population can take preventative measures as well. Traditional public service announcements and flyers educate the public about the risk of standing water and ways in which the public can reduce risk of mosquito-transmitted infections, but these efforts are rarely targeted to the neighborhoods and times when education will have the largest impact. We can change this by making timely risk data available to the public and communicated in easy to understand terms. Ultimately, this will lead to a safer and healthier Chicago, at limited taxpayer expense.






## Objectives




































Goal Type	Business Objective	SMARRT Deliverable	Success Criteria
<b>Deliver Predictive Models</b>	City of Chicago can perform strategic and targeted intervention activities by identifying areas which have high risk of mosquitos carrying West Nile Virus	Deliver predictive models using mosquito activity data, weather data and data from the city of Chicago like zoning, demographic, income characteristics etc.  Stretch Goal: Evaluation of factors which affect spraying effectiveness.	Models evaluated on common classification and regression metrics using cross-validation and hold out datasets.
<b>Deliver Real Time Actionable Insights</b>	City of Chicago and it's residents will have access to up to date predictions of West Nile Virus threat levels	Deliver a dashboard with real time updated threat levels, predictions of outbreaks, identification of high-risk regions like hospitals,	Voice of customer feedback score on usability and value of dashboard or mobile

	among other valuable insights	playgrounds, senior living facilities etc. Weekly or monthly reports by zip-code or neighborhood can be generated.	application.
--	-------------------------------	--	--------------

## Data Sources

To develop the predictive models and insightful dashboards, the team will scrape data from various sources from the internet. As we have shown in the table below, data needed to address this problem are varied in size, complexity, variety, quality and availability. After a preliminary investigation into these data, the team has assigned qualitative scores to help determine feasibility for model building.

	Size	Complexity	Variety	Quality	Availability
	Unknown	Unknown	Unknown	Unknown	Unknown
	<100MB	No preprocessing required for consumption	Numerical	...	Some risk in availability
	100-500MB	Some preprocessing required for consumption	Numerical + Categorical	Quality, not vetted	Partially Available
	500-1GB	Substantial preprocessing & preparation required	Temporal + Numerical + Categorical	...	Available, not vetted
	1GB+	Substantial & complex data munging required	Spacio-temporal + Numerical + Categorical	High Quality & vetted	Available & vetted

Source	Dataset	Description	Size	Complexity	Variety	Quality	Availability
Chicago Department of Public Health	Mosquito trap and West Nile Virus, 2007-2018	27000 records over 11 years with location, mosquito species, and presence of west nile virus					
National Oceanic and Atmospheric Administration	Daily weather data, 2007-2018	Daily precipitation and high/low/average temperature readings for Chicago and surrounding areas					
Webscraping or Satellite Imagery Analysis	Geospatial Water Body Information	Locations and metadata of water bodies and marshlands for Chicago and surrounding areas					
Unknown	Aviary data	Bird population and death rates by location for Chicago and surrounding areas					
United States Census Bureau's American Community Survey	Sociodemographic data, 2007-2018	Poverty rates, socioeconomic status, education status, unemployment status over 11 years					
Cook County Data Portal	Geospatial Hospital & School Locations	Locations of hospitals, schools, and senior assisted living facilities to characterize areas of highly vulnerable populations					
Unknown	Financial Data	Financial impact of West Nile Virus: estimated per-infection treatment costs, spraying and prevention costs, impact on businesses and local economy					

1. The primary dataset to be used for predictive modeling , the Mosquito trap and West Nile Virus test data, 2007-2018, can be obtained from Chicago Department of Public Health (CDPH) via the [Chicago](#)

[Data Portal](#). These data provide mosquito trap locations (latitude & longitude), species specific mosquito counts and West Nile virus test results. This dataset is small: roughly 27,000 observations for 12 variables.

2. Daily weather data will be obtained from the [National Oceanic and Atmospheric Administration](#) (NOAA), which is part of the United States Department of Commerce. Due to the partial Federal government shutdown from January 26th 2019, the NOAA data are currently unavailable. Alternate sources are being identified from [Kaggle](#) or scraping public datasets on [wunderground](#).
3. Geospatial predictors such as locations of bodies of water & marshland will be considered. A feasibility study is required to determine if obtaining such information is possible within the scope of this project. Webscraping or satellite imagery analysis are two options to identify areas with a higher concentration of bodies of water and standing water.
4. Although aviary data for specific bird populations would be very useful, these data are not available in any known dataset.
5. To assess potential human impact of West Nile virus outbreaks in mosquitos, we will investigate sociodemographic data from the [United States Census Bureau's American Community Survey](#) (ACS) obtained from <ftp2.census.gov>. We will extract measures of poverty, low socioeconomic status, and vulnerability to West Nile virus, particularly neuroinvasive disease, such as older age. We will use 5-year ACS summary data at the Census block group level. CDPH mosquito traps are already geocoded to provide latitude & longitude, but we will further perform spatial joins to identify the Census block group and Chicago neighborhood community area in which each trap is found.
6. We will attempt to obtain locations of hospitals, schools, senior assisted living facilities and other areas with high concentrations of vulnerable populations. Hospital and school locations were obtained from the [Cook County Data Portal](#). These data locations (latitude & longitude) will also be spatially joined to obtain Census block group and Chicago neighborhood community area. All spatial joins will be performed using Census TIGER/Line files (i.e. GIS shapefiles). We will compute driving time (minutes), public transit time (minutes) and distance in feet as-the-crow-flies from each Census block group centroid to the nearest hospital.
7. To assess the financial impact of preventing West Nile virus infections, we will use published cost analyses that have assessed the burden of medical care for West Nile virus patients. These data will be used to estimate financial impact of West Nile virus given the predicted mosquito count, likelihood of West Nile virus being present, and demographics of the human population nearby. We will also use published analyses of the cost of mosquito abatement programs for West Nile virus.

# Deliverables

---

## Dashboard Visualization

Maps are a powerful tool that can be used to effectively visualize geospatial data. SMARRT Consulting group will deliver several maps to support the findings from our analysis. The maps will be incorporated into different Tableau dashboards that will be interactive. This will enable the CDPH to do some self-service analytics within the environment we create.

The main type of map we plan to deliver are heat maps. Heat maps are a graphical representation of data that transform the quantitative data into color. Typically, these colors are on a spectrum that is ranging from red to yellow to green, with the different color gradients representing a different numeric value. The main benefit to using heat maps is the ability to quickly ingest a large amount of data. Looking at a list of Chicago neighborhoods or zip codes and the number of West Nile virus infections in each area would be a lot of data for the naked eye to comprehend. However, looking at that same data on a map that is divided into Chicago's neighborhoods or zip codes and color coded according to the number of West Nile virus infections, is a completely different story. On the map, it will be very easy to immediately identify pockets where the number of West Nile virus infections are very high (in red) or very low (in green). Combining the color dimension with the geospatial will allow very simple ease of use for the CDPH.

Other aspects we plan to incorporate into our dashboards are markers to indicate high risk areas. We are defining high risk areas as those where a large outbreak of West Nile virus infections would have the worst impact due to a more sensitive population group. This would include areas with a lot of schools or daycare centers or areas with retirement homes.

One way this can be done is using the interactive hover feature that is built into Tableau. Following along the same example as earlier, if you were to hover your mouse over a certain neighborhood or zip code that had a high number of West Nile virus infections and was colored red, a small text box would appear and give you additional information on that particular area. This information could include the population in that area, whether a spraying was done, the number of schools and daycares, the number of retirement homes, etc. There are several options for what types of supplemental data can be provided in these text boxes.

Another option for indicating these high-risk areas would be to overlay small pictorials in the areas. For instance, a small clipart picture of a school in school zones.

SMARRT Consulting group is dedicated to working closely with the CDPH to deliver the solution that would be most beneficial.

## Mobile Application

According to the 2017 Mary Meeker report, the number of hours spent on the internet is still increasing each year, but the split between desktop and mobile is becoming more and more pronounced. In 2016, Americans were spending 3+ hours per day on mobile (that's 10 times more than in 2008) and just 2.2 hours per day on a desktop or laptop (no change since 2008).<sup>4</sup>

Given these recent trends, SMARRT Consulting Group will deliver a dashboard for both desktop and mobile platforms. The initial version of the mobile application will be limited, but the goal is to supply the CDPH with

---

<sup>4</sup> 25 Mobile App Usage Statistics To Know In 2019. (n.d.). Retrieved from <https://mindsea.com/app-stats>



one version for both desktop and mobile users. The desktop dashboard will be the primary channel for consuming the data with the mobile application providing a “lite” version of the dashboard. Like the desktop dashboard, the mobile application will allow users to interact with the data through filters which re-render the data visuals, but the scope of the mobile application is to focus on the classification of the West Nile virus and identify potential high-risk areas for contracting the West Nile virus.

SMARRT Consulting Group will monitor the mobile application usage, and if the demand for the mobile application is high then a full version of the mobile application will be delivered as part of a future update. The mobile application will be developed in the R Shiny library and deployed to the web through the Shiny Server Open Source.

## Predictive Modeling

Based on a literature review<sup>5</sup> of West Nile virus epidemiology<sup>6</sup> and predictive modeling, several factors have been identified that could be used to predict the outbreaks of the West Nile virus. The spread of mosquito populations for certain species which carry the virus has been proven to be a strong predictor of outbreaks.<sup>7</sup> Human infection with West Nile virus occurs primarily from mosquito vectors and is not transmitted human-to-human. Birds also carry the virus and function as a reservoir such that ongoing bird-to-mosquito and mosquito-to-bird transmission keeps the virus active. Surveillance of birds has been used successfully to identify potential West Nile virus outbreaks in humans<sup>8</sup>, but this is impossible without active aviary surveillance data which is currently unavailable in Chicago. Weather patterns are important in affecting mosquito populations. Bodies of water such as rivers, lakes and marshland have been shown to correlate with mosquito populations, as has standing water in swimming pools and abandoned properties.

We aim to use geospatial, temporal, and observational data to predict mosquito populations as well as mosquito infections with the West Nile virus. To model the mosquito populations, traditional time series forecasting models like ARIMA, ETS, STLF or TSLM can be investigated. More sophisticated methods like Prophet may prove useful in cases of changing seasonality or trend patterns and presence of outliers. Other methods proven useful for predicting spread of Influenza cases called Method of Analogues<sup>9</sup> will also be investigated.

For regression and classification parts of the problem, traditional statistical approaches like Lasso regression, Principal Components Regression and Multinomial Regression will be investigated. Machine learning

---

<sup>5</sup> Amy V. Bode, James J. Sejvar, W. John Pape, Grant L. Campbell, Anthony A. Marfin; West Nile Virus Disease: A Descriptive Study of 228 Patients Hospitalized in a 4-County Region of Colorado in 2003, *Clinical Infectious Diseases*, Volume 42, Issue 9, 1 May 2006, Pages 1234–1240

<sup>6</sup> Calistri P, Giovannini A, Hubalek Z, et al. Epidemiology of west nile in europe and in the mediterranean basin. *Open Virol J*. 2010;4:29-37. Published 2010 Apr 22

<sup>7</sup> Kilpatrick AM, Pape WJ. Predicting human West Nile virus infections with mosquito surveillance data. *Am J Epidemiol*. 2013;178(5):829-35.

<sup>8</sup> Eidson M, Kramer L, Stone W, Hagiwara Y, Schmit K, New York State West Nile Virus Avian Surveillance Team. Dead bird surveillance as an early warning system for West Nile virus. *Emerg Infect Dis*. 2001;7(4):631-5.

<sup>9</sup> Cécile Viboud, Pierre-Yves Boëlle, Fabrice Carrat, Alain-Jacques Valleron, Antoine Flahault; Prediction of the Spread of Influenza Epidemics by the Method of Analogues, *American Journal of Epidemiology*, Volume 158, Issue 10, 15 November 2003, Pages 996–1006

approaches like RandomForest, Xgboost models and deep learning networks may prove useful if the data have high interactions and nonlinearities.

Feature reduction using Lasso Regression, Variable Importance estimation in RandomForest or embedding matrices from Autoencoder networks may be used to identify useful predictors in case the data suffer from multicollinearity or large dimensions.

We will use holdout sets to evaluate model performance on un-seen data, while using cross-validation and validation datasets to perform model parameter selection and hyperparameter tuning.

## Risks

---

The team will be spending a significant portion of the time in the front end of the project compiling, cleansing and preparing these data for further consumption. However, as shown above, the sources of some of the expected important predictive variables can prove to be challenging to obtain within our project's timeline. The team will strive to strike a balance between developing complex yet accurate models with many predictors and keeping a low risk towards delivery of our end product to the city.

For the team's stretch goal of determining the factors which affect spraying effectiveness, spraying data is only available for 2 of the 11 years of mosquito data. This can substantially curb how well we can determine what the city of Chicago can do to improve spraying effectiveness. Furthermore, the team has investigated and learned that the city did perform regular spraying campaigns throughout the 11-year period - even though those data are not available. This results in a latent confounded variable for modeling team: how can the impacts of such sprayings be accounted for? How much error could they introduce?

## Software & Analytical Tools

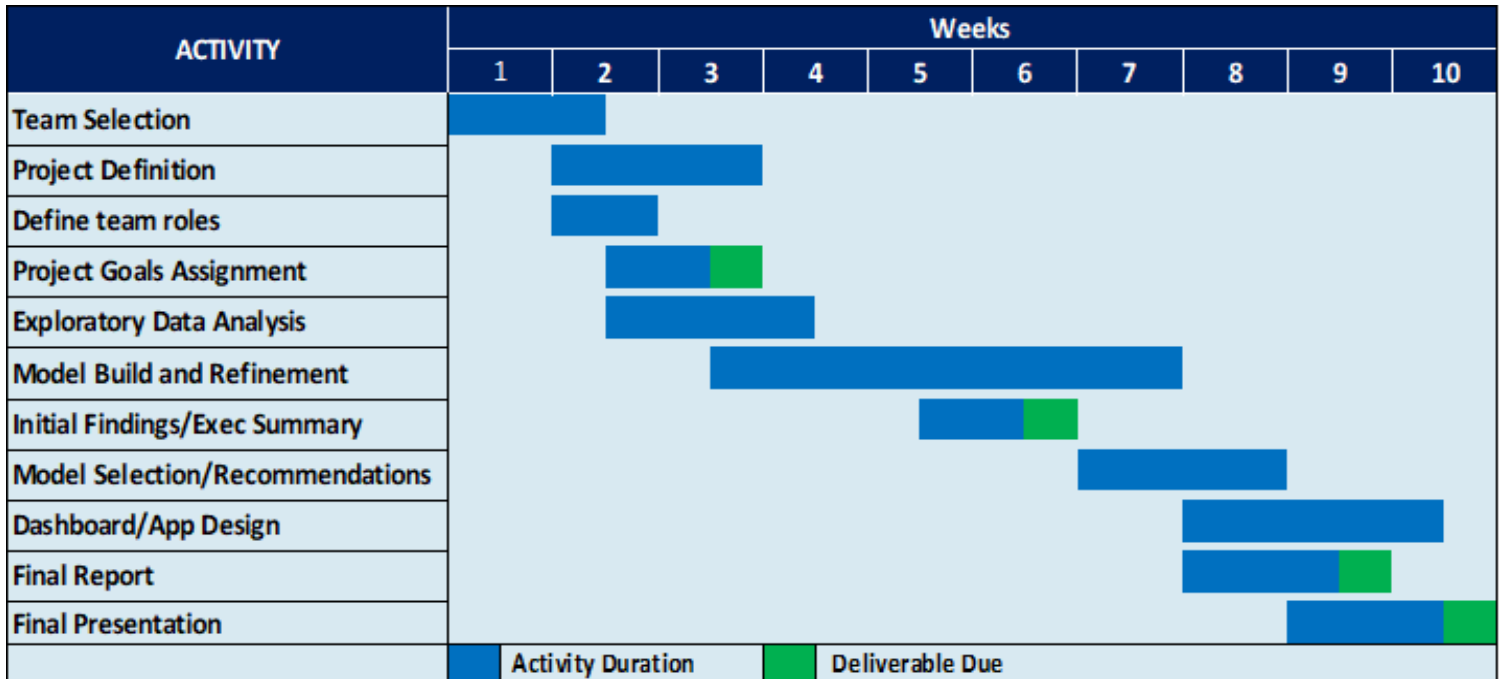
---

SMART Analytics will use the following toolkit to complete the project:

- CRAN R 3.3+
- Python 3.5
- RStudio
- Jupyter Lab
- Tableau Desktop
- Microsoft Excel
- Microsoft PowerBI
- SAS JMP Pro 13

# Project Plan

The project plan for this proposal is outlined below. This project is set to accomplish 11 major activities and four milestones for this project within its ten-week duration. To accomplish this ambitious timeline the team will work on their respective tasks individually while coordinating with other team members on data preparation, modeling, and documentation of project progress. The team will share findings and next steps during weekly all team meetings. We will be leveraging a management approach that is consistent with the CRISP-DM process model. The CRISP-DM model is flexible and can be customized easily, as the process outlines the steps involved in performing data science activities from business goals to deployment while indicates how iterative this process is. Any deviations from the timeline below will be communicated to the project sponsor in the weekly status reports.



**Team Selection** - Recruit the appropriate team for the project with similar interests and complementary skills.

**Project Definition** - Identify the data set, business scenario, analytical direction.

**Define Team Roles** - Identify team strengths and interests to best support the project and delineate the scope of work for each role.

**Project Goals** - Establish the business case, proposed analysis, project timeline, and outcomes.

**Exploratory Data Analysis** - Conduct a preliminary analysis of the data to gain insights.

**Model Build and Refinement** - Develop and improve various models to predict the potential risk of West Nile prevalence in Chicago.

**Initial Findings and Executive Summary** - Deliver a report stating the problem, the approach taken, data analysis and preliminary conclusions.

**Model Selection and Recommendations** - The models will be evaluated via a variety of performance metrics as well as interpretability to derive recommendations to the client as well as begin dashboard and app design.

**Dashboard and App Design** - Deliver a functional Dashboard and a preliminary beta application.

**Final Report** - Compose a comprehensive report explaining our work including our analysis, conclusions, and business recommendations.

**Final Presentation** - Deliver the final report in an online presentation, illustrating our problem, our analysis, takeaways and client recommendations. The App and Dashboard will be showcased via a demonstration.

## The Team

---

The SMARRT consulting group is a diverse team consisting of professionals from a range of educational backgrounds with a wealth of practical experience. Our team comes from a variety of industries including: R&D, academia, finance, automotive, operations, and analytics. We know that this team and all the skills that we bring to the table will be able to exceed your expectations on this project.

**Andrew Cooper** has a Master of Public Health in epidemiology and has 15+ years of work experience as a statistical analyst and manager of a software development team that developed web-based data entry, data management and analytic tools. In addition to contributing to analytic design, he will be involved with identifying data sources; obtaining, cleaning and reshaping data; and building predictive models.

**Rachel Dudle** is a new addition to the SMARRT Consulting Group. While she has only 4+ years' experience, she has established herself with strong skills in building visualizations and dashboards. Her experience ranges from the Financial sector, to manufacturing to pharmaceuticals.

**Stephen Hage** has a diverse career background, having worked in operations, sales, marketing and analytics. His strengths as a modeler and data storyteller will help this project mature from concept to effective product. He will be a bit involved with most aspects but will also be the sales and marketing lead for SMARRT Consulting Group.

**Ted Inciong** has an M.S. in Information Technology from the Illinois Institute of Technology and has 10+ years of experience in the Financial sector.

**Mike Kapelinski** is a Project manager that has a M.S. in Biotechnology and 9+ years working in a variety of fields across science such as clinical oncology, pharmaceutical manufacture, and R&D. He brings domain knowledge of the sciences and experience leading a variety of projects to help this team deliver goals on time and above expectations.

**Rahul Sangole** has 11+ years of work experience in the Automotive industry, primarily in Engineering, Quality and Analytics, leveraging both predictive modeling and six sigma to drive organizational changes using analytics.